International Journal of Computer Science Engineering
and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Vol. 4, Issue 3, Jun 2014, 83-92


# A COMPARATIVE ANALYSIS OF THE PERFORMANCE OF CLOUD COMPUTING WITH JAVA AND HADOOP

**DALYA RAAD ABBAS[1] & SANJAY T. SINGH[2]**

[1]M. Sc. Student, Department of CS & IT, SHIATS, Allahabad, Uttar Pradesh, India

Department of Computer Science, Ministry of Higher Education and Scientific Research, College of Science, Kirkuk University, Kirkuk, Iraq

[2]Assistant Professor, SHIATS, Allahabad, Uttar Pradesh, India

## ABSTRACT

The software applications of the cloud computing caused accumulation in data, managing and distributing these data in an effective way and standard time had become an urgent need, hadoop is an open source software platform, had been developed by Apache to deal with these massive data using parallel processing.

In this paper both cloud computing and hadoop had been covered, constrained of both strong and weak points separately, the main goal of this paper is to merge both hadoop and cloud computing, address and analyze the performance of both when they work together, based on this tiny virtual cloud built, hadoop with single cluster installed, two scenarios proposed to examine and analyze the environment contained both cloud computing and hadoop, to make this analyzing near to the reality, the first case constrained of managing data with the tools of hadoop as a standard, the last case had been worked only with the cloud, these cases customized to be able to manage and distribute files of different kinds numeric or strings, these files of different sizes, the results had been showed that processing files of huge sizes had took very long time exceed one hour without hadoop, while processing these files with hadoop and cloud had showed significant improvement, the built environment which contain both hadoop and cloud computing, recorded high performance reached to 90%.

**KEYWORDS:** Cloud Computing, Hadoop, Java

## INTRODUCTION

Based on "National institute standards and technology"(NIST) in 2013[10] the Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

[8]Hadoop is an open source framework, which deal with the distributed system and has the ability to analysis and execute BIG DATA or we can say files that contain huge size ( terabytes and more).

The main characteristics of hadoop are, that hadoop can run in clusters of machines either these machines be in the same data center or it can be by the “cloud”, the beauty of hadoop that, it’s based with the ability to handle the hardware failure in the cluster, so it’s already has the ability to handle any unexpected failure, hadoop has the ability to analysis and execute the data that has no fix size, also can deal with the grown data, easily by adding more nodes to the cluster, allows the

users of hadoop to write codes which can be consider as a simple codes comparing with the huge size of data that can analysis and deal with it.

The basic components of hadoop(five daemons), Name Node, Data Node, Secondary Name Node, Job Tracker, Task Tracker, the first three parts come under HDFS, the HDFS can support the very big sizes of data terabytes and more of that, also it has the ability to spilt these files of data and spread them in the cluster of hadoop, and to any number of machines, and HDFS has the ability to save the files by provides the replication term, each portion of data that had been stored in a machine had the duplicate of it in two or three other machines, this can provide the reliability to the stored data and saving these data from being losing, the HDFS has the ability to serve any number of incoming requests from the clients and in the same time in fast (this feature gives more popularity of the Hadoop), HDFS can deal with the increasing number of the nodes in efficient manner and spread the partition data among them.

The two remain parts come under Map Reduce the heart of hadoop, Map Reduce is a programming framework tool, originally created by Google but later had been developed by hadoop, Map Reduce used to connect and for reducing the execution time and get the required result in hadoop, Map Reduce can work with multiple kind of data, relational, structural, un structural, Map Reduce can work to improve their goals with Big Data, so we can see HDFS and Map Reduce both have common key working with Big Data

## RELATED WORKS

- The researchers in this paper used the hadoop for connecting the RDF(resource description framework) that used by semantic web technologies which work with cloud computing, the semantic web technologies worked as a toll for handling data intensive application, the researchers build a framework for store the huge number of RDF data in the HDFS, also they had present an algorithm for generating the best plan for SPARQL protocol and RDF query language by using Map Reduce framework.

The results had showed that the suggested service integrity assurance framework, is more efficient and has the ability to detect the malicious slave in the cloud computing environment with the use of hadoop Map Reduce.

- In this paper the researchers used the hadoop and cloud with the help of multiplication matrix to build cloud that cover the full crime information in Texas city, they used Joynet Smart Data Center, they installed Flex Cloud for cloud environment from the Joynet Smart Data Center, they used the matrix multiplication to divided the Texas city map with the help of Matrix multiplication of (4096 * 4096) size for each matrix, after that they installed hadoop Map Reduce for spilt, multiplication (keys), amongst mappers(tasks) then these mappers will be summed in the reduce task, then with the installing of varies size hadoop clusters (1,2,4,8,16,32) virtual machines .

The Texas city divided to 32 area, so every information entry about rime will be attached the area number, according to the matrix multiplication matrix that implement in the hadoop of different cluster size, the comparison result showed that the hadoop with 32 virtual machine improved their speed, also in the another comparison of the time used in each cluster of hadoop, it showed that the 32 virtual machine hadoop cluster take the lee time to get the results.

- In this paper the researchers focused on making a study of the implementation on data mining without and with hadoop cloud computing, first they had implement the Apriroi algorithm without MapReduce programming language, then they had implemented with the MapRreduce programming language, their strategy had been tested

using IBM's Quest Synthetic Data Generator, they had used 100000 transaction and each transaction contain 5 data set of items, he had built a cluster contain 10 pc's with different hardware requirement.

In the first part of this paper a rich material will be produce on the Cloud Computing and Hadoop, to understand the characteristics, architecture and the work nature of both Cloud Computing and Hadoop separately, The second part of this paper is listing a group of the previous works that had done in the previous years, also studying these works with some kind of details, The third part of this paper list the objectives of this paper, the forth part aims to inauguration a tiny virtual cloud, then inauguration of Hadoop single cluster in one machine, building two different Cases to investigate the behavior of the work of Cloud Computing and Hadoop, The fifth part will presents the results of processing of 72 files of different sizes and contents from the two Cases that had been built in the fourth part, then a discussion is obtained with results, to determine if there is an improvement, based on the merging of the Cloud Computing with Hadoop, The sixth part present a final conclusion.

## OBJECTIVE

The objective of this paper to build a tiny virtual cloud and merge hadoop with the cloud, then built two different kinds of cases one for the hadoop and the cloud both together, the other case is the cloud itself, then test the performance of hadoop and cloud, test the results to see if the performance of hadoop and the cloud can be increased or not, meanwhile test the results of the cloud without hadoop, in the last a comparison built to analysis the obtained results.

## METHODOLOGY

After finishing the installation of single cluster hadoop in one machine, test if the hadoop is working probably or not.

Now the hadoop is ready, the health of the HDFS can be monitor, by using this site, and connecting with HDFS with this port number (50070).

Also monitoring the performance of Map Reduce, and monitoring every request come from the client, and how this request will be served, how many time it will take till finishing each request, from this port number this job can be accomplish (50030).

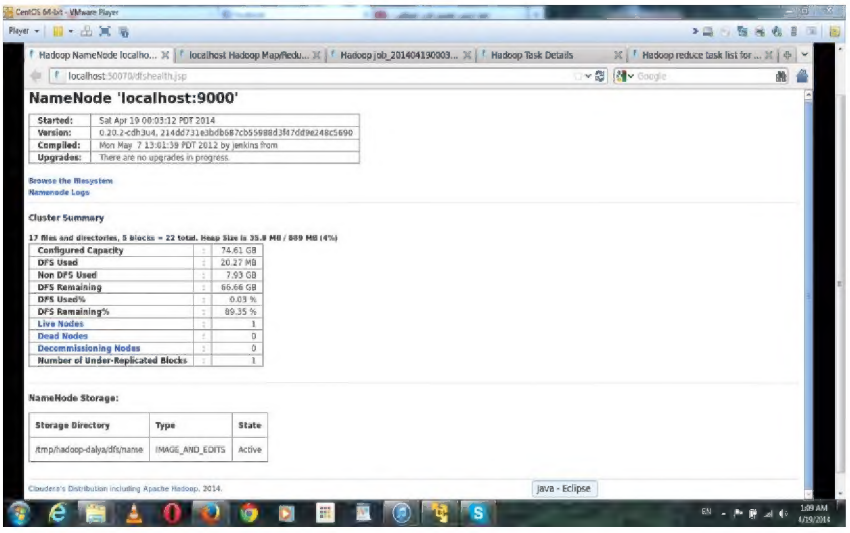


**Figure 1: The HDFS of Hadoop**

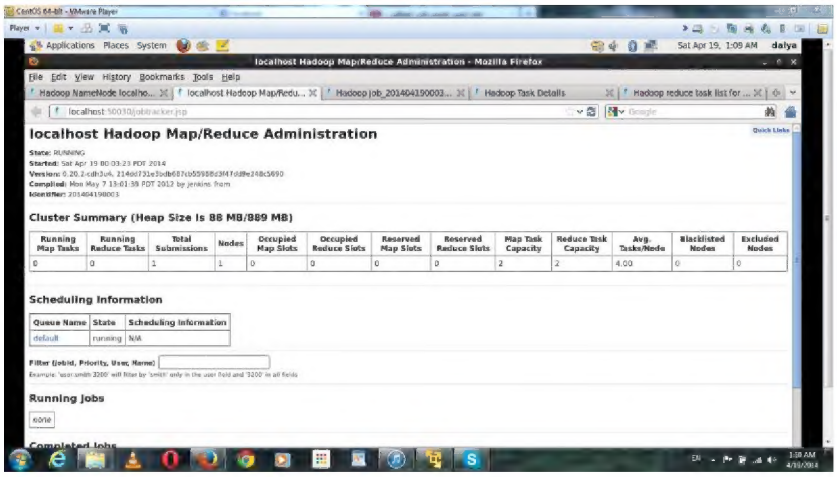


**Figure 2: The Map Reduce of Hadoop**

Since hadoop had been built with Java language, also the second case of the cloud using only java, so java programming language had been also installed, the platform used to work with java, and connecting java with hadoop was Eclipse, one of the platforms used to run java language, what make this platform special, it's characterized with simplicity and with friendly interface.

Two different kinds of scenarios had been built to test the performance of hadoop and cloud computing, the first case is to test the performance of hadoop with cloud computing, the second case is to test the performance of cloud computing itself, without hadoop, the performance of this two cases aims on the ability to process the big data or files of big sizes.

**Case (1)**

In this case had been created four of Map Reduce programs, using the basic tools in Hadoop, to test the performance of the normal hadoop in two ways.

The first one is building a MapReduce program to test the performance of cloud computing and hadoop in processing the files of words type.

The second one is build a MapReduce program, to process big sizes files of numbers contents only.

The MapReduce programs used in this case are:

- Search Program.
- Maximum Program.

**Search Program**

The purpose of this program is for searching for a specific content, also print how many this content repeated in a specific file, the entries to this program must be only of alphabetic contents, to ensure processing of files with this program done in an effective way, first of all particular job created to this program, the job retrieves the files needed to be processed from the server where its already reside on it, also to processing it in parts:

- **Partition Part:** In this part, the job request from the client to give the word he/she want to find, then the job start creating key and value, then the job broke the contents of file, taking each word and assign a key for each word in

the file, each broken word come under value and each value take a particular numeric key, the job produce as output set of values and its keys with the word the client want to find.

- **Finding Word Part:** The job take the intermediate outputs, and the word the client want to find as an input to the next part of the program, in this part the job takes the needed word as an searching key, start reading the intermediate words one after one, whenever the word the client want to find is matched then the job count this word, till reaching to the end of these intermediate outputs, then the founded word and how many times it found print out to a file in the mentioned directory in the server as the client request.

**Maximum Program**

This program is for processing files and to extract the maximum number from the whole content of file, and produce the result in the file print to the server, the file type that this program uses is only numeric files, to ensure processing of files with this program done in an effective way, when the client request to retrieve the maximum number of numeric file, job create and request from the client to locate in the server which needed to find the maximum number of it also to processing it in parts*:*

- **Parition Part:** In this part, the job breaks the contents of file to set of values, each value take a numeric key, the numbers treated as a strings and comes under the attribute value, and each value has a unique key, the job produce the output set of values and its keys and prepare it to the next part of finding Maximum number.
- **Finding the Maximum Number Part:** The job take the intermediate outputs, then start searching to the minimum number, then the job take this minimum number as a maximum number and start compare the whole intermediate values with this number whenever found a number that is greater than this number take it as a maximum, till reach the end of the intermediate number, then the job print the maximum number in a file and print it to the directory located from the client.

**Case (2)**

In this case the test based on testing the behavior of cloud computing without using hadoop, and to make the test based on the consistent basis, programs used in this testing are the same in the name but of course completely different from the programs used in the last two cases, also since the tested behavior based on testing the behavior of the cloud using files of numbers, also using files of characters, so in the third case, had been created two programs, one of the two programs created to processing the different sizes of files, containing only numbers, the second program for processing different sizes of alphabetic files.

The Map Reduce Programs are:

- Search Program.
- Maximum Program.

**Search Program**

The purpose of this program is for searching for a specific content, also print how many this content repeated in a specific file, the entries to this program must be only of alphabetic contents, using normal java, had been created normal program for search complete file and find the particular word, also counting how many this word repeated in this program.

### Maximum Program

This program is for processing files and to extract the maximum number from the whole content of file, and produce the result in the file print to the server, this program work only with numeric content of files, using normal java, had been created normal program for finding the maximum number in the complete file, find the minimum number in the file, then taking this number as a the maximum number and making comparing with all the content of file, then print out the maximum found number.

## RESULTS

According to the two cases had been created in the cloud computing environment, to make the simulation near as much as possible to the reality, in this paper had been created different files of txt type, To test the virtual cloud and its ability to process the two cases, so in this paper had been used different size of files, to make the analysis of cloud computing and hadoop, fair as much as possible, the size of the files used in the cloud, twelve different size of files, these files are contain only numbers, and their sizes are (30, 60, 90, 120, 150, 180, 200, 210, 240, 270, 300, 400) MB, also another twelve different size of files, but these files contain only words, and their sizes are (30, 60, 90, 120, 150, 180, 200, 210, 240, 270, 300, 400) MB.

### Case (1)

Using 24 files in total of txt files, two different programs had been tested in this case and under different circumstances the results that had been obtained from the first case using normal hadoop.

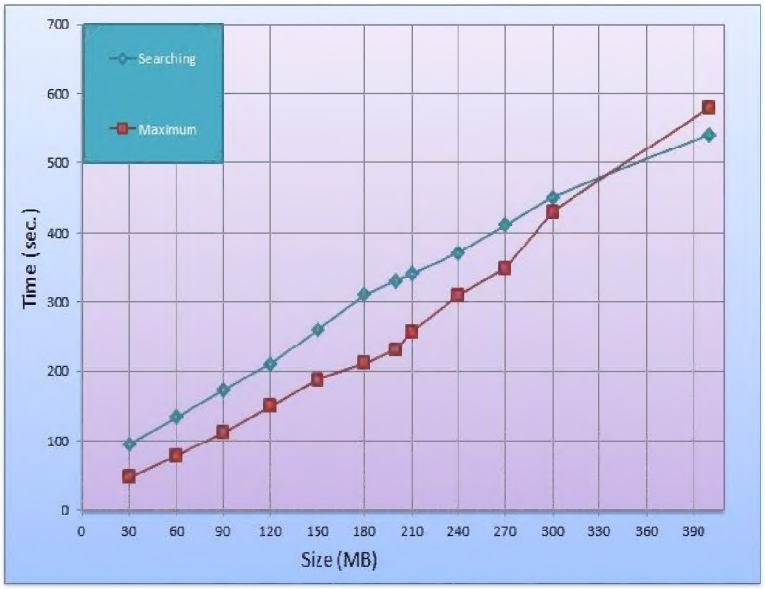


**Figure 3: The Case (1) Results**

### Case (2)

In the second and last case the using only normal java without using hadoop, and using in total 24 files of txt types the results was :

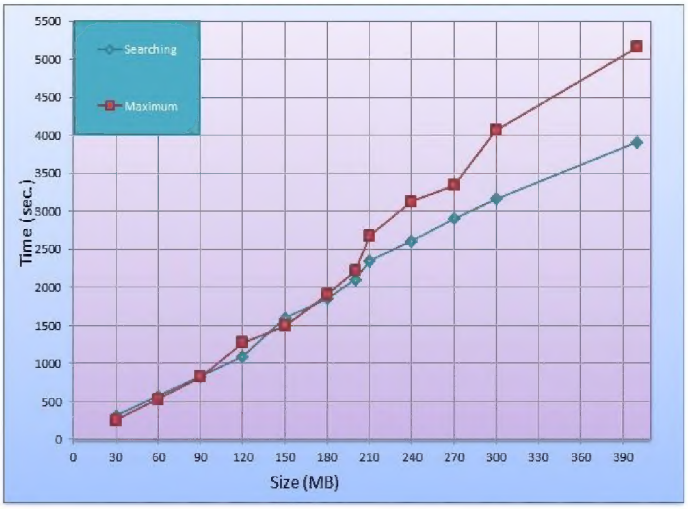


**Figure 4: The Case (2) Results**

## DISSCUSIONS

### Maximum Program

The behavior of this program had been test using case(1) in the virtual cloud with basic tools of hadoop, in the case(2) in the virtual cloud without using hadoop, the results of the processing time in these three cases using the same txt files and the same sizes of these files are shown in the table below:

Maximum program to get the maximum number in the file, for the minimum file of size 30 MB was (47.442) seconds, and for the maximum file size of 400 MB was (579.802), that's mean in the 30 MB file the elapsed time in minutes was 79 seconds mean less than one minute and as a maximum in the 400 MB was 9 minutes and 66 seconds, also the results taken with different circumstances.

According to the above table, case(1) that had been tested under the virtual cloud with the use of the basic tool of hadoop had proved better improvement comparing with the same program running under only virtual cloud without using hadoop case (2) the improvement range was between (81.55005658 %) as a minimum and (90.41391053 %) as a maximum, so the total average improvement of case (1) was (87.99167%).

### Search Program

This program had been tested with three cases, in the case (1) this program had been tested with the virtual cloud with hadoop that uses here its basic tools, the second case the program had been tested with case (2) using the virtual cloud without using hadoop, the processing time that the program taking with 12 txt files with these three cases are:

**Table 1: Maximum Program Comparison**

| Size | Case(1) | Case(2) | Improvement % |
|---|---|---|---|
| 30 | 95.344 | 310.107 | 69.25448 |
| 60 | 134.373 | 567.461 | 76.32031 |
| 90 | 173.133 | 830.795 | 79.16056 |
| 120 | 210.033 | 1087.827 | 80.69243 |
| 150 | 260.212 | 1600.527 | 83.7421 |
| 180 | 310.068 | 1850.783 | 83.24666 |
| 200 | 330.807 | 2100.779 | 84.25313 |
| 210 | 340.018 | 2350.445 | 85.53389 |

**Table 1: Contd.,**

| | | | |
|---|---|---|---|
| 240 | 370.319 | 2611.097 | 85.81749 |
| 270 | 411.182 | 2905.583 | 85.84855 |
| 300 | 450.681 | 3162.019 | 85.74705 |
| 400 | 541.083 | 3911.686 | 86.16752 |

In the search program was only for searching a specific word chosen randomly, as a result the elapsed time for execution these 12 txt files, as shown in the above figure, was for the minimum size of file that had been 95.344 seconds, and for the maximum file size of 400 MB was (541.083) seconds, so the execution of the file of 30 MB takes one minute and 58 seconds exactly, and in the maximum file of 400 MB the elapsed time was 9 minutes and one millisecond, these results was taken different circumstances .

According to the above table, the behavior of search program had been compared, the behavior of search program in case (1) had been compared with the search program in case (2), so based on case (2) it shows that the processing time in the case (1) had been showed clearly improvement comparing with case(2), the range of this improvement was as a minimum (69.25448%) and the maximum improvement was (86.16752%).

**Table 2: Search Program Comparison**

| Size | Case(1) | Case(2) | Improvement % |
|---|---|---|---|
| 30 | 47.442 | 257.139 | 81.55005658 |
| 60 | 77.57 | 531.842 | 85.41484125 |
| 90 | 111.625 | 825.239 | 86.47361552 |
| 120 | 150.272 | 1272.391 | 88.18979386 |
| 150 | 188.021 | 1497.148 | 87.44138856 |
| 180 | 211.516 | 1906.738 | 88.90691852 |
| 200 | 230.979 | 2219.084 | 89.59124576 |
| 210 | 257.06 | 2681.594 | 90.41391053 |
| 240 | 309.485 | 3128.155 | 90.10646851 |
| 270 | 347.861 | 3344.469 | 89.5989169 |
| 300 | 430.17 | 4071.34 | 89.4341912 |
| 400 | 579.802 | 5166.962 | 88.77866723 |

## CONCLUSIONS

In this paper the behavior of the cloud with hadoop had been tested, each and every case and each and every file had been processed with different circumstances, to test the ability of the cloud to deal the different requests that come from different clients, with the normal hardware requirement that shown in the table below, the cloud and hadoop in case (1) improved its ability to take the different requests come from the clients, serve these requests with an optimum time, without affecting the machine health that hold the hole cluster, the maximum time was taken with cloud and hadoop was approximately 10 minutes with files of 400 MB.

**Table 3: The Hardware and Software Requirement**

| | Hardware and Software Requirements | |
|---|---|---|
| 1. | Hard Disk | 80 GB SCSI |
| 2. | Memory | 2 GB |
| 3. | Processor | Intel(R)Core 1-24430M |
| 4. | CPU | 2.40 GHz |
| 5. | Operating system | SentOS final 64 bit |

Also the behavior of the cloud alone without hadoop had been tested, dealing with different request come from different clients, had been showed that serving each request from the clients, had made the machine acting very slowly and the processing time with the increasing of file size had also increased, which mean the cloud act in very bad way with the increasing of the number of the incoming requests from the clients, and also with the increasing of the file sizes.

## REFERENCES


1. M. Husain, L. Khan, M. Kantarciogluz, B. Thuraisinghamx, "Data Intensive Query Processing for Large RDF Graphs Using Cloud Computing Tools "IEEE 3rd International Conference on Cloud Computin, in 2010 pp.1 – 10.
2. Y. Ren, W. Tang, T. Varvarigou, "A SERVICE INTEGRITY ASSURANCE FRAMEWORK FOR CLOUD COMPUTING BASED ON MAPREDUCE"Proceedings of IEEE CCIS, in 2012 pp.240 –244.
3. W. Luo, N. Golpavar,C. Cardenas, A. Chronopoulos "Benchmarking Joyent SmartDataCenter for Hadoop MapReduce and MPI Operations ", in IEEE.
4. Y. Bao, L. Ren, L. Zhang, X. Zhang, Y. Luo, "Massive Sensor Data Management Framework in Cloud Manufacturing Based on Hadoop," in IEEE, 2012, pp. 397 – 401.
5. Cloud Computing: A Practical Approach by Anthony T. Velte, Toby J. Velte and Robert Elsenpeter (2010).
6. Hadoop For Dummies®, Special Edition by Robert D. Schneider (2012).
7. Hadoop in Practice by Alex Holmes (2012).
8. Hadoop: The Definitive Guide 3rd edition by Tom White (2012).
9. Hadoop in Action by Chuck Lam (2011).
10. NIST(National Institute of Standerds and Technology ),The NIST Cloud Computing Standreds Roadmap.America:Special Publication 800-146,by Lee Badger, Tim Grance, Robert Patt-Corner and Jeff Voas (2012).
11. Welcome to Apache Hadoop, 2010. http://hadoop.apache.org/
12. Download Eclipse for linux, censtos, http://www.if-not-true-then-false.com/2010/linux-install-eclipse-on-fedora-centos-red-hat-rhel/.
13. Download Java JDK for centos http://www.oracle.com/technetwork/java/javase/downloads/jre7-downloads-1880261.html.
14. Download VMPlayer https://my.vmware.com/web/vmware/downloads.
15. Downloiad CentOS http://isoredirect.centos.org/centos/6/isos/x86_64/.
16. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
17. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.